Package 'FMAT'

August 11, 2023

Title The Fill-Mask Association Test

Version 2023.8

Date 2023-08-08

Maintainer Han-Wu-Shuang Bao <baohws@foxmail.com>

Description The Fill-Mask Association Test ('FMAT') is an integrative, versatile, and probability-based method that uses Masked Language Models to measure conceptual associations or relations (e.g., attitudes, biases, stereotypes, social norms, cultural values) as propositional representations in natural language. The supported language models include 'BERT' (Devlin et al., 2018) <arXiv:1810.04805> and its model variants available at 'Hugging Face' <https://huggingface.co/models?pipeline_tag=fill-mask>. 'Python' ('conda') environment and the 'transformers' module can be installed automatically using the FMAT_load() function. Methodological references and technical details are provided at <https://psychbruce.github.io/FMAT/>.

License GPL-3

Encoding UTF-8

URL https://psychbruce.github.io/FMAT/

BugReports https://github.com/psychbruce/FMAT/issues

SystemRequirements Python (>= 3.6.0)

Depends R (>= 4.0.0)

Imports PsychWordVec, psych, reticulate, text, data.table, stringr, forcats, glue, cli, purrr, plyr, dplyr, tidyr

Suggests bruceR, nlme, parallel

RoxygenNote 7.2.3

NeedsCompilation no

Author Han-Wu-Shuang Bao [aut, cre] (<https://orcid.org/0000-0003-3043-710X>)

Repository CRAN

Date/Publication 2023-08-11 08:20:02 UTC

R topics documented:

. *A simple function equivalent to* `list`.

Description

A simple function equivalent to `list`.

Usage

```
.(...)
```

Arguments

...	Named objects (usually character vectors for this package).

Value

A list of named objects.

Examples

```
.(Male=cc("he, his"), Female=cc("she, her"))
list(Male=cc("he, his"), Female=cc("she, her"))  # the same
```

FMAT_load *Initialize running environment and (down)load language models.*

Description

Initialize running environment and (down)load language models.

Usage

```
FMAT_load(models)
```

Arguments

models	Language model names (usually the BERT-based models) at HuggingFace. For a full list of available models, see https://huggingface.co/models?pipeline_tag=fill-mask&library=transformers

Value

A named list of fill-mask pipelines obtained from the models. The returned object *cannot* be saved as any RData. You will need to *rerun* this function if you restart the R session.

All downloaded models are saved at your local folder "C:/Users/[YourUserName]/.cache/".

See Also

PsychWordVec::text_init

FMAT_query

FMAT_query_bind

FMAT_run

Examples

```
models = FMAT_load(c("bert-base-uncased", "bert-base-cased"))
```

FMAT_query *Prepare a data.table of queries and variables for the FMAT.*

Description

Prepare a data.table of queries and variables for the FMAT.

Usage

```
FMAT_query(
  query = "Text with [MASK], optionally with {TARGET} and/or {ATTRIB}.",
  MASK = .(),
  TARGET = .(),
  ATTRIB = .(),
  unmask.id = 1
)
```

Arguments

query	Query text (should be a character string/vector with at least one [MASK] token). Multiple queries share the same set of MASK, TARGET, and ATTRIB. For multiple queries with different MASK, TARGET, and/or ATTRIB, please use FMAT_query_bind to combine them.
MASK	A named list of [MASK] target words. Must be single words in the vocabulary of a certain masked language model. For model vocabulary, see, e.g., https://huggingface.co/bert-base-uncased/raw/main/vocab.txt Note that infrequent words may be not included in a model's vocabulary, and in this case you may insert the words into the context by specifying either TARGET or ATTRIB.
TARGET, ATTRIB	A named list of Target/Attribute words or phrases. If specified, then query must contain {TARGET} and/or {ATTRIB} (in all uppercase and in braces) to be replaced by the words/phrases.
unmask.id	If there are multiple [MASK] in query, this argument will be used to determine which one is to be unmasked. Defaults to the 1st [MASK].

Value

A data.table of queries and variables.

See Also

FMAT_load

FMAT_query_bind

FMAT_run

Examples

```
FMAT_query("[MASK] is a nurse.", MASK = .(Male="He", Female="She"))

FMAT_query(
  c("[MASK] is {TARGET}.", "[MASK] works as {TARGET}."),
  MASK = .(Male="He", Female="She"),
  TARGET = .(Occupation=cc("a doctor, a nurse, an artist"))
)

FMAT_query(
  "The [MASK] {ATTRIB}.",
  MASK = .(Male=cc("man, boy"),
           Female=cc("woman, girl")),
  ATTRIB = .(Masc=cc("is masculine, has a masculine personality"),
             Femi=cc("is feminine, has a feminine personality"))
)

FMAT_query(
  "The association between {TARGET} and {ATTRIB} is [MASK].",
  MASK = .(H="strong", L="weak"),
  TARGET = .(Flower=cc("rose, iris, lily"),
             Insect=cc("ant, cockroach, spider")),
  ATTRIB = .(Pos=cc("health, happiness, love, peace"),
             Neg=cc("death, sickness, hatred, disaster"))
)
```

FMAT_query_bind *Combine multiple query data.tables and renumber query ids.*

Description

Combine multiple query data.tables and renumber query ids.

Usage

```
FMAT_query_bind(...)
```

Arguments

...	Query data.tables returned from FMAT_query.

Value

A data.table of queries and variables.

See Also

FMAT_load

FMAT_query

FMAT_run

Examples

```
FMAT_query_bind(
  FMAT_query(
    "[MASK] is {TARGET}.",
    MASK = .(Male="He", Female="She"),
    TARGET = .(Occupation=cc("a doctor, a nurse, an artist"))
  ),
  FMAT_query(
    "[MASK] occupation is {TARGET}.",
    MASK = .(Male="His", Female="Her"),
    TARGET = .(Occupation=cc("doctor, nurse, artist"))
  )
)
```

FMAT_run *Run the fill-mask pipeline on multiple models.*

Description

Run the fill-mask pipeline on multiple models.

Usage

```
FMAT_run(
  models,
  data,
  file = NULL,
  progress = c(FALSE, TRUE, "none", "text", "time"),
  parallel = FALSE,
  ncores = 4,
  warning = TRUE
)
```

Arguments

models	Language model(s): • Model names (usually the BERT-based models) at HuggingFace. • A list of mask filling pipelines loaded by FMAT_load. * You will need to **rerun** FMAT_load if you **restart** the R session.
data	A data.table returned from FMAT_query or FMAT_query_bind.

file	File name of .RData to save the returned data.
progress	Show a progress bar: "none" (FALSE), "text" (TRUE), "time".
parallel	Parallel processing (NOT suggested). Defaults to FALSE. If TRUE, then models must be model names rather than from FMAT_load. * For small-scale data, parallel processing would instead be *slower* because it takes time to create a parallel cluster.
ncores	Number of CPU cores to be used in parallel processing.
warning	Warning of out-of-vocabulary word(s). Defaults to TRUE.

Details

The function will also automatically adjust for the compatibility of tokens used in certain models: (1) for uncased models (e.g., ALBERT), it turns tokens to lowercase; (2) for models that use <mask> rather than [MASK], it automatically uses the corrected mask token; (3) for models that require a prefix to estimate whole words than subwords (e.g., ALBERT, RoBERTa), it adds a certain prefix (usually a white space; \u2581 for ALBERT and XLM-RoBERTa, \u0120 for RoBERTa and DistilRoBERTa).

Note that these changes only affect the token variable in the returned data, but will not affect the M_word variable.z Thus, users may analyze their data based on the unchanged M_word rather than the token.

Value

A data.table (of new class fmat) appending data with these new variables:

- model: model name.
- output: complete sentence output with unmasked token.
- token: actual token to be filled in the blank mask (a note "out-of-vocabulary" will be added if the original word is not found in the model vocabulary).
- prob: (raw) conditional probability of the unmasked token given the provided context, estimated by the masked language model.

  * It is NOT SUGGESTED to directly interpret the raw probabilities because the *contrast* between a pair of probabilities is more interpretable. See summary.fmat.

See Also

FMAT_load

FMAT_query

FMAT_query_bind

summary.fmat

Examples

```
# Running the example requires the models downloaded
# You will need to rerun `FMAT_load` if you restart the R session

models = FMAT_load(c("bert-base-uncased", "bert-base-cased"))

query1 = FMAT_query(
  c("[MASK] is {TARGET}.", "[MASK] works as {TARGET}."),
  MASK = .(Male="He", Female="She"),
  TARGET = .(Occupation=cc("a doctor, a nurse, an artist"))
)
data1 = FMAT_run(models, query1)
summary(data1, target.pair=FALSE)

query2 = FMAT_query(
  "The [MASK] {ATTRIB}.",
  MASK = .(Male=cc("man, boy"),
           Female=cc("woman, girl")),
  ATTRIB = .(Masc=cc("is masculine, has a masculine personality"),
             Femi=cc("is feminine, has a feminine personality"))
)
data2 = FMAT_run(models, query2)
summary(data2, mask.pair=FALSE)
summary(data2)

query3 = FMAT_query(
  "The association between {TARGET} and {ATTRIB} is [MASK].",
  MASK = .(H="strong", L="weak"),
  TARGET = .(Flower=cc("rose, iris, lily"),
             Insect=cc("ant, cockroach, spider")),
  ATTRIB = .(Pos=cc("health, happiness, love, peace"),
             Neg=cc("death, sickness, hatred, disaster"))
)
data3 = FMAT_run(models, query3)
summary(data3, attrib.pair=FALSE)
summary(data3)
```

`LPR_reliability` *Reliability analysis (Cronbach's α) of LPR.*

Description

Reliability analysis (Cronbach's α) of LPR.

Usage

```
LPR_reliability(fmat, item = c("query", "T_word", "A_word"), by = NULL)
```

Arguments

fmat	A data.table returned from summary.fmat.
item	Reliability of multiple "query" (default), "T_word", or "A_word".
by	Variable(s) to split data by. Options can be "model", "TARGET", "ATTRIB", or any combination of them.

Value

A data.table of Cronbach's α.

---

summary.fmat *[S3 method] Summarize the results for the FMAT.*

Description

Summarize the results of *Log Probability Ratio* (LPR), which indicates the *relative* (vs. *absolute*) association between concepts.

The LPR of just one contrast (e.g., only between a pair of attributes) may *not* be sufficient for a proper interpretation of the results, and may further require a second contrast (e.g., between a pair of targets).

Users are suggested to use linear mixed models (with the R packages nlme or lme4/lmerTest) to perform the formal analyses and hypothesis tests based on the LPR.

Usage

```
## S3 method for class 'fmat'
summary(
  object,
  mask.pair = TRUE,
  target.pair = TRUE,
  attrib.pair = TRUE,
  warning = TRUE,
  ...
)
```

Arguments

object	A data.table (of new class fmat) returned from FMAT_run.
mask.pair, target.pair, attrib.pair	Pairwise contrast of [MASK], TARGET, ATTRIB? Defaults to TRUE.
warning	Warning of out-of-vocabulary word(s). Defaults to TRUE.
...	Other arguments (currently not used).

Value

A data.table of the summarized results with Log Probability Ratio (LPR).

See Also

FMAT_run

Examples

```
# see examples in `FMAT_run`
```

Index